Hybrid Spintronic-CMOS Spiking Neural Network with On-Chip Learning: Devices, Circuits, and Systems

Abhronil Sengupta,* Aparajita Banerjee, and Kaushik Roy

School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907, USA

(Received 25 June 2016; revised manuscript received 28 September 2016; published 8 December 2016)

Over the past decade, spiking neural networks (SNNs) have emerged as one of the popular architectures to emulate the brain. In SNNs, information is temporally encoded and communication between neurons is accomplished by means of spikes. In such networks, spike-timing-dependent plasticity mechanisms require the online programing of synapses based on the temporal information of spikes transmitted by spiking neurons. In this work, we propose a spintronic synapse with decoupled spike-transmission and programing-current paths. The spintronic synapse consists of a ferromagnet–heavy-metal heterostructure where the programing current through the heavy metal generates spin-orbit torque to modulate the device conductance. Low programing energy and fast programing times demonstrate the efficacy of the proposed device as a nanoelectronic synapse. We perform a simulation study based on an experimentally benchmarked device-simulation framework to demonstrate the interfacing of such spintronic synapses with CMOS neurons and learning circuits operating in the transistor subthreshold region to form a network of spiking neurons that can be utilized for pattern-recognition problems.

DOI: 10.1103/PhysRevApplied.6.064003

I. INTRODUCTION

Brain-inspired computing models have emerged as one of the most powerful tools for pattern-recognition and classification problems over the past few decades [1]. Such schemes attempt to develop abstract models of the communication and functionalities involved in the neurons and the synapses in the human brain in order to construct computing tools efficient at recognition and cognitive tasks. However, implementation of such non–von Neumann computing schemes on general-purpose supercomputers have not been able to harness the energy efficiency of the human brain. The sequential fetch, decode, and execute cycles involved in traditional von Neumann computing are in complete contrast to the parallel, event-driven processing involved in the mammalian cortex. For instance, the IBM Blue Brain project [2] utilized the Blue Gene supercomputer to simulate brain activity in animals and consumed orders of magnitude more energy than the brain, even at neuron firing rates much slower than the biological time scale.

Custom CMOS analog and digital VLSI neurocomputing platforms have been also utilized to implement neuron and synapse functionalities. BrainScaleS [3], SpiNNaker [4], and IBM TrueNorth [5] are instances of such neurocomputers based on conventional CMOS technology. However, the significant mismatch between the neuroscience mechanisms involved in the brain and the CMOS transistors have limited the capability of such computing technologies to achieve the area or power efficiency of the brain. For example, four 8-T static random-access memory (SRAM) cells (32 CMOS transistors) are required to implement the functionality of a single 4-bit synapse in a digital CMOS implementation [6].

Recently, neurocomputing architectures based on emerging post-CMOS technologies have gained popularity, as they offer a direct mapping to many of the neuroscience mechanisms involved in biological synapses [7–11] and neurons [12–14]. In order to achieve an integration density similar to the brain's, neuromorphic-computing architectures aim to achieve a FAN-OUT of 10 000 for each neuron, thereby requiring orders of magnitude more synapses than neurons. Additionally, unsupervised learning using spike-timing-dependent plasticity (STDP), or other Hebbian learning rules, requires online programing of synapses during spike transmission. Hence, a nanoelectronic device emulating synaptic functionalities is an essential component of spiking neuromorphic architectures.

In this work, we propose a ferromagnet–(FM–)heavy-metal (HM) multilayer structure where spin-orbit torque induced by the programing current flowing through the HM is the main underlying physical mechanism for generating synaptic plasticity. The ferromagnet is part of a magnetic-tunneling-junction (MTJ) structure where spike voltage transmitted through the MTJ gets modulated by the MTJ conductance. The proposed three-terminal device structure offers the advantage of decoupled spike-transmission and programing-current paths, thereby leading to an efficient implementation of on-chip learning. Furthermore, the proposed synapse can be programed at low current magnitudes and small programing-time durations, and it thereby consumes orders of magnitude lower programing energy in comparison to other state-of-the-art emerging synaptic

*asengup@purdue.edu

devices. We discuss a comprehensive framework for simulating such spintronic synapse-based spiking neural systems from the device (including calibration to experimental results) to the system level for performing recognition tasks.

II. SPIKING NEURAL NETWORKS: PRELIMINARIES

A. Neuron and synapse dynamics in spiking neural networks

A synapse is a junction connecting two neurons. The transmitting neuron is termed as the preneuron, while the receiving neuron is termed as the postneuron. The preneuron transmits a train of voltage spikes which may be represented by a set of Dirac-δ functions at the time instants t_f,

$$V_{\text{pre}} = \sum_f \delta(t - t_f). \tag{1}$$

The synapse response to such a spike train is modeled by

$$\tau_{\text{post}} \frac{dI_{\text{post}}}{dt} = -I_{\text{post}} + w \sum_f \delta(t - t_f), \tag{2}$$

where I_{post} is the postsynaptic current produced by the synapse characterized by weight w, and τ_{post} is the time constant of the postsynaptic current. Hence, the postsynaptic current increases by an amount modulated by the synapse conductance (the weight) at each spike instant and then starts decaying exponentially. The temporal dynamics of the leaky integrate-and-fire neuron in response to such a postsynaptic current is given by

$$\tau \frac{dV_{\text{mem}}}{dt} = -V_{\text{mem}} + R_{\text{mem}} \sum_i I_{\text{post},i}, \tag{3}$$

where V_{mem} is the membrane potential, R_{mem} is the membrane resistance, $I_{\text{post},i}$ is the postsynaptic-current input from the ith neuron, and τ is the membrane time constant. Figure 1 shows the temporal characteristics of the neuron and the synapse in response to a series of voltage spikes transmitted from the preneuron. When the neuron's membrane potential V_{mem} crosses the threshold V_{thres}, the membrane potential gets reset to V_{reset} and does not vary for a time duration termed as the refractory period.

B. Learning: STDP

According to the theory of Hebbian learning [15], synaptic weight or conductance is modulated depending on the spiking patterns of the preneuron and the postneuron. STDP, a form of Hebbian learning, states that the weight of the synapse increases (decreases) if the preneuron spikes before (after) the postneuron. Intuitively, this

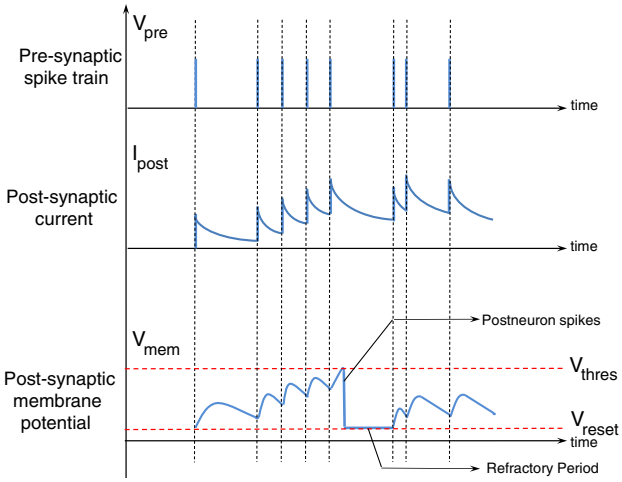

FIG. 1. Neuron and synapse dynamics in response to a spike train.

postulate signifies that the synapse strength should increase if the preneuron spikes before the postneuron, as the preneuron and the postneuron appear to be temporally correlated. The relative change in synaptic strength decreases exponentially with the timing difference between the preneuron and postneuron spikes. The STDP characteristics have been formulated in a mathematical framework based on measurements for rat hippocampal glutamatergic synapses [16],

$$\begin{aligned} \Delta w &= A_+ \exp\left(\frac{-\Delta t}{\tau_+}\right), \qquad \Delta t > 0 \\ &= -A_- \exp\left(\frac{\Delta t}{\tau_-}\right), \qquad \Delta t < 0. \end{aligned} \tag{4}$$

Here, A_+, A_-, τ_+, and τ_- are constants and $\Delta t = t_{\text{post}} - t_{\text{pre}}$, where t_{pre} and t_{post} are the time instants of pre- and postsynaptic firings, respectively. We refer to the case of $\Delta t > 0$ ($\Delta t < 0$) as the positive (negative) time window for learning.

C. Spike-frequency adaptation

In order to model spike-frequency-adaptation mechanisms observed in biological neurons, an additional slowly varying adaptation parameter a is introduced in the temporal dynamics of the neuron as

$$\tau \frac{V_{\text{mem}}}{dt} = -V_{\text{mem}}(1 + a) + R_{\text{mem}} \sum_i I_{\text{post},i}. \tag{5}$$

The adaptation parameter a increases every time the neuron spikes; otherwise, it decays exponentially. This model implies that, in a case where a neuron starts spiking at a high frequency, the leak parameter starts to increase to reduce its spike frequency.

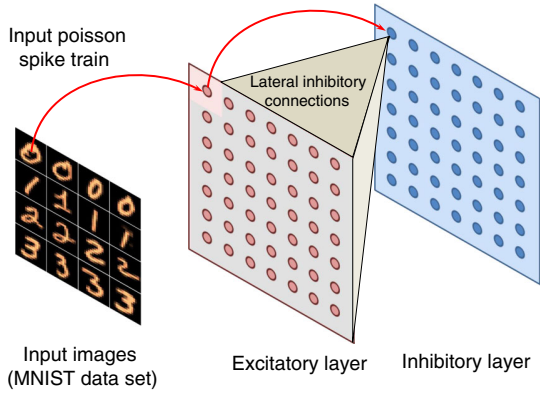

FIG. 2. Network connectivity utilized for pattern recognition. Neurons with lateral inhibitory connections receive input Poisson spike trains with an average rate proportional to the pixel intensity.

D. Network connectivity

Figure 2 shows the network connectivity of spiking neurons utilized for pattern-recognition problems. Such a network topology has been shown to be efficient in several pattern-recognition problems, such as digit recognition [17] and sparse encoding [18]. The input image pixels are encoded as Poisson spike trains with an average rate directly proportional to the pixel intensity. These input spike trains are received by all neurons in an excitatory layer through synapses whose weights are learned using STDP. Each neuron in the excitatory layer is connected to a corresponding neuron in an inhibitory layer such that a spike in the excitatory neuron triggers a spike in the corresponding neuron in the inhibitory layer. Each neuron in the inhibitory layer is connected to all neurons in the excitatory layer except for the neuron from which it received the input. This connectivity helps us to implement lateral inhibitory connections in the excitatory layer, such that when one neuron starts to spike in response to some input pattern, it prohibits the other neurons from spiking. However, in order to prevent a particular neuron from dominating the spiking pattern due to lateral inhibitory connections, a spike-frequency-adaptation mechanism is also implemented in each neuron. The neurons in the excitatory layer are assigned classes based on their highest response (spike frequency) to input training patterns.

III. SPINTRONIC SYNAPSE

A. Spin-orbit torque-driven motion of Dzyaloshinskii domain walls

In this section, we provide a brief discussion on the underlying physical phenomena involved in current-induced domain-wall motion in HM-FM-insulator (I) multilayer structures.

Recent experiments on magnetic nanostrips of Pt/CoFe/MgO and Ta/CoFe/MgO have revealed high domain-wall velocities due to charge-current densities that are 2 orders of magnitude lower than that achievable by conventional spin-transfer torque (STT) [19]. Additionally, domain-wall motion is observed to be against the direction of electron flow (i.e., in the direction of current flow) in multilayer structures with Pt as the underlayer, thereby suggesting that current-induced spin-orbit torque is the main mechanism of domain-wall motion in such multilayer structures (with a negligible contribution from conventional STT) [19]. In such magnetic heterostructures with high perpendicular magnetocrystalline anisotropy, spin-orbit coupling and broken inversion symmetry leads to the stabilization of homochiral domain walls through the Dzyaloshinskii-Moriya exchange interaction (DMI) [20]. We restrict our analysis for Pt/CoFe/MgO multilayer structures in this work due to the possibility of achieving high domain-wall velocities (approximately 400 m/s) [21–23]. However, the analysis can be easily extended to other magnetic heterostructures with different underlayers.

Such an interfacial DMI at the FM-HM interface leads to the formation of a Néel domain wall with left-handed chirality for Pt/CoFe/MgO multilayer structures [19,21–23]. The DMI strength in such structures with HM underlayers has been observed to be sufficiently strong to impose a Néel-wall configuration in FMs where conventional magnetostatics would have yielded a Bloch configuration [19]. When an in-plane charge current is injected through the HM, a transverse spin current is generated due to the deflection of opposite spin polarizations on the top and bottom surfaces of the HM. This phenomenon is termed as the spin Hall effect [24] and arises as a consequence of spin-orbit torque. The accumulated spins at the FM-HM interface lead to DMI-stabilized Néel-domain-wall motion. The direction of domain-wall motion is in the direction of charge-current flow, and the final magnetization of the ferromagnet is given by the cross product of the direction of the injected spins at the FM-HM interface and the magnetization direction of the FM at the domain-wall location.

B. Device proposal for spintronic synapse

Such spin-orbit torque-driven domain-wall motion in FMs due to charge-current flow through a HM underlayer leads to the possibility of a device structure that can manifest decoupled spike-transmission (read) and programing-current (write) paths. We propose a three-terminal device structure consisting of a FM lying on top of a HM (Fig. 3). The FM is part of a MTJ structure where the FM is separated from a pinned layer (a magnetic region whose magnetization is fixed) by a tunneling-oxide barrier (MgO). The FM has two additional pinned layers on either side to ensure that the domain wall stabilizes at the extreme locations of the FM for sufficiently large values of the programing current. While the spike current flows through the MTJ structure between terminals $T1$ and $T3$, the

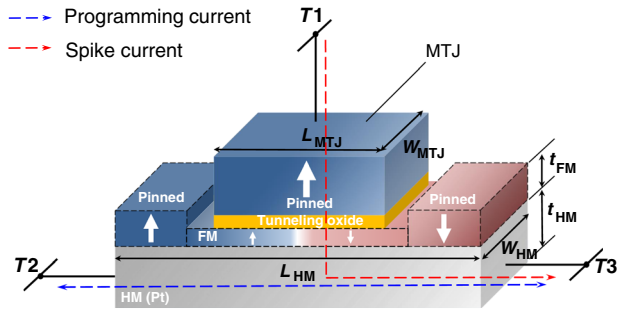

FIG. 3. Device structure for a spintronic synapse with decoupled spike-transmission and programing-current paths. Spike current flows through the MTJ structure between terminals $T1$ and $T3$. Programing current flows through the HM between terminals $T2$ and $T3$.

programing current flows through the HM layer between terminals $T2$ and $T3$. Note that a preliminary synaptic-device proposal based on Bloch-domain-wall motion due to spin-orbit torque was explored previously in Ref. [25]. However, an external magnetic field is required to modulate the device conductance during learning. Furthermore, the magnet width is not scalable beyond 100 nm to ensure Bloch-wall orientation. The current device proposal based on Néel-wall motion is not only more energy efficient but requires no external magnetic field for domain-wall motion due to the inherent interfacial DMI. Furthermore, this work provides a synergistic device-circuit-system perspective for the implementation of STDP in spiking neural networks (SNNs) utilizing the proposed spintronic device as the core building block.

The location of the domain wall in the FM encodes the resistance of the device lying in the path of the spike current between terminals $T1$ and $T3$ and thereby implements the synaptic functionality. On the other hand, the programing-current path is completely decoupled (between terminals $T2$ and $T3$) and the resistance in the path of the programing current is mainly determined by the HM resistance. It is worth noting here that, although some amount of spike current will flow through the HM, the magnitude of this current can be maintained to sufficiently low values below the domain-wall depinning current since the synapses are required to drive CMOS neurons operating in the subthreshold regime.

C. Synaptic-plasticity mechanism

Programing current flowing from terminal $T2$ to terminal $T3$ results in domain-wall motion in the same direction, so the $+z$ domain in the FM starts to expand, and vice versa. For a given duration of the programing-current pulse, the domain-wall displacement is directly proportional to the magnitude of the programing current.

On the other hand, the device conductance between terminals $T1$ and $T3$ varies linearly with the domain-wall position. Let us denote the conductance of the device when the entire FM magnetization is parallel (antiparallel) to the Pinned layer as G_P (G_{AP}); i.e., the domain wall is at the extreme right (left) of the FM. Thus, for an intermediate position of the domain wall at a position x from the left edge of the MTJ, the device conductance between terminals $T1$ and $T3$ is given by

$$G_{eq} = G_P \frac{x}{L} + G_{AP} \cdot \left(1 - \frac{x}{L}\right) + G_{DW}, \tag{6}$$

where L denotes the length of the MTJ excluding the domain-wall width and G_{DW} represents the conductance of the wall region. It is worth noting here that L, G_{DW}, G_P and G_{AP} are all constants (for a constant voltage drop across the MTJ). Owing to such a linear relationship between the domain-wall position and the device conductance, the programing current is directly proportional to the change in device conductance (which encodes the synaptic weight) for a fixed duration of the programing signal.

D. Spiking neuromorphic architecture based on spintronic synapse

Figure 4 represents a possible arrangement of a spintronic synapse with access transistors $M_{A1} - M_{A4}$ to decouple the programing- and spike-current paths. The access transistors act as switches for selecting the appropriate terminals of operation for the device. The operating mode of the synapse—i.e., the spike-transmission mode or the programing mode—is accomplished by the control signal POST. The POST signal is activated during the programing mode of operation of the synapse.

The PRE line is used to pass the necessary amount of programing current required for the corresponding weight change involved due to the delay between the preneuron and postneuron spikes. A negative (positive) current should flow through the HM for the negative (positive) time window duration. Since the programing-current amplitude is directly proportional to the amount of weight change, the current signal flowing through the HM should vary in a similar fashion as the STDP learning curve (exponentially) with the time delay between the preneuron and postneuron spikes.

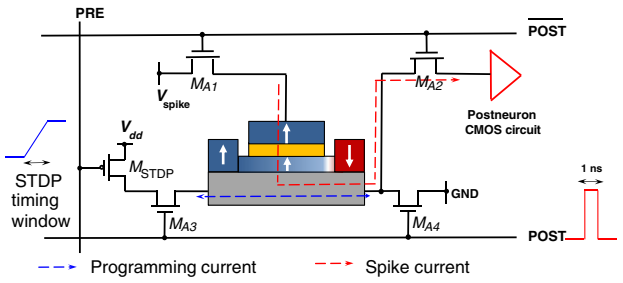

FIG. 4. Spintronic synapse with access transistors to decouple the programing- and spike-current paths.

For simplicity, let us discuss the case for the positive time window. The exponential variation of current through the HM can be obtained by a transistor operating in the subthreshold regime since the current flowing through the transistor will vary exponentially with the gate-to-source voltage. Thus, for a linear increase of voltage of the PRE line with time, the transistor M_{STDP} will be driven from the cutoff to the saturation regime when the POST signal is activated, and an appropriate programing current should flow through the HM. It is worth noting here that the HM resistance equals approximately a few hundred ohms and the maximum programing current required is approximately a few tens of μA, thereby leading to a very small voltage drop across the device when the POST signal is activated. Figure 4 shows the interface circuits involved in the synapse programing for the positive time window. A similar approach can be adopted to program the synapses for the negative time window (by utilizing an NMOS operating in subthreshold saturation driven by a linearly increasing gate voltage to pass the programing current from terminal $T3$ to terminal $T2$) and the two learning circuits for the negative and positive timing windows have to be activated sequentially every time the preneuron spikes. Since the time duration involved in programing is approximately a few nanoseconds—in comparison to learning time constants used in this work of, approximately, microseconds—the POST signal essentially samples the necessary amount of programing current from the PRE line (the programing-current magnitude determined by the M_{STDP} transistor).

In our proposed programing scheme, we program the synapses only when the postneuron spikes. Hence, in order to account for the negative and positive time windows involved in STDP learning, the POST signal should be activated with a delay corresponding to the time duration of the negative timing window in order to sample the programing-current contributions from the learning circuits for both of the timing windows.

An arrangement of synapses in an array fashion (as shown in Fig. 5), interfaced with CMOS neurons, can lead to dense spiking neuromorphic architectures. Please note that the access transistors M_{A2} and M_{A4} for terminal $T3$ of the device (Fig. 4) can be shared across the row such that the corresponding horizontal line connecting terminals $T3$ for the devices in a particular row are driven to ground (the POST signal is high) or the postneuron circuit (the POST signal is low). Details of the CMOS circuits involved in the programing scheme and neuron implementation are discussed in Sec. IV.

IV. CMOS LEARNING AND NEURON CIRCUITS

A. Subthreshold circuit for STDP learning

The circuit involved in generating the PRE signal is discussed in this section. Figure 6 shows the subthreshold CMOS circuit used to generate the PRE signal for preneuron A connecting to postneurons C and D. We discuss the mechanism for generating the signal for the positive time window. A similar design can be used to generate the programing current for the negative time window. The circuit was originally proposed in [26] as a reset and discharge synapse. However, it failed to emulate the postsynaptic dynamics of biological synapses, as the circuit response depends only on the previous input spike [27]. In this work, we employ this circuit to implement STDP learning in our proposed device.

The transistor M_p acts as a switch. When the positive time window starts, the transistor M_p receives a low-active pulse and gets turned *on*. As a result, the node PRE, A is set to the bias voltage V_w. After the transistor M_p is switched *off*, the transistor M_t, operating in the subthreshold saturation regime, provides a constant current to linearly charge the capacitor C_p at a rate I_t/C_p. Hence, if the transistor

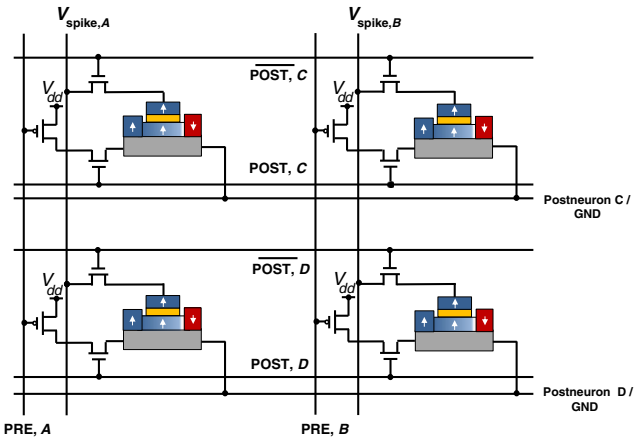

FIG. 5. Possible arrangement of synapses in an array interfaced with CMOS neurons and programing circuits. Shown are synapses connecting preneurons A and B to postneurons C and D.

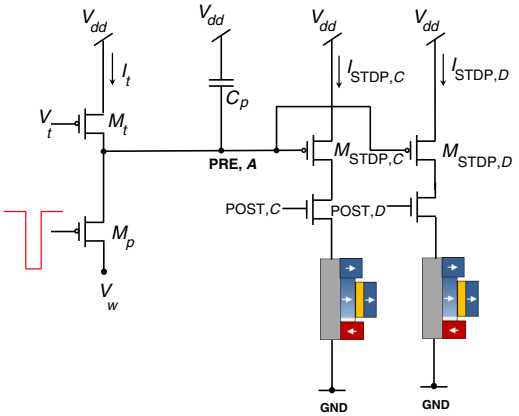

FIG. 6. Subthreshold CMOS circuit utilized for generating the programing current involved in STDP learning (the circuit for the positive time window shown) for preneuron A, connecting to postneurons C and D.

M_{STDP} is operated in subthreshold saturation, exponential dynamics will be observed in the output current I_{STDP}. The current flowing through transistor M_{STDP} for an input pulse at time $t = t_n$ is given by (if the POST signal is active)

$$I_{\mathrm{STDP}} = I_0 e^{\frac{-U_T C_p (t-t_n)}{k I_t}}, \tag{7}$$

where k is the subthreshold slope factor and U_T is the thermal voltage. Hence, whenever the preneuron spikes, the circuits for generating the STDP characteristics for the negative and positive time windows are activated sequentially. When learning starts for the positive timing window, a short pulse is applied to the gate of the transistor M_p so that the circuit is reset and the node PRE, A is charged to V_w. When the postneuron does not spike, the transistor M_{STDP} is in cutoff since the POST signal is deactivated and the access transistors for programing are turned *off*. Once the postneuron spikes, the programing-current path gets activated and the transistor M_{STDP} switches to the subthreshold saturation regime and transmits the necessary amount of programing current through the device. Note that apart from the transistor M_{STDP} (one transistor for each of the positive and negative timing windows), the entire learning circuitry can be shared across the column of the crossbar array.

The operation is discussed in detail in Fig. 7. Let us first describe the case for the positive timing window, i.e., postneuron spiking after the preneuron [Fig. 7(a)]. $-\Delta$ ($+\Delta$) represents the duration during which the learning circuit for the negative (positive) timing window is activated sequentially for the corresponding preneuronal firing event. The control signal POST is activated after a duration (Δ) when the postneuron spikes. As described in the figure, magnitude of the programing pulse is determined by the current being passed by the programing transistor M_{STDP} (the value of the PRE voltage when the POST signal is active), and the duration is determined by the duration of the POST signal. Since the PRE signal varies in an approximately microsecond time scale and almost does not change during the programing-time duration (an approximately nanosecond time scale), it ensures that the programing-current magnitude is almost constant and is equal to the sampled value from the exponential STDP dynamics corresponding to the appropriate spike-timing difference. As mentioned previously, since the programing current magnitude is directly proportional to the amount of change in the MTJ conductance, exponential STDP characteristics are implemented in the spintronic device. Similar discussions are valid for the negative timing window [Fig. 7(b)] where the postneuron spikes before the preneuron. In this case, the POST signal is activated during the negative window ($-\Delta$) and the NMOS transistor passes an appropriate amount of programing current in the opposite direction through the device. Circuit-level simulations confirming the proposal are demonstrated in Fig. 11(b).

B. Differential-pair-integrator circuit for postsynaptic current generation

The differential-pair-integrator (DPI) circuit has been a popular mechanism for generating synaptic dynamics [28], and integration of such DPI circuits with memristor synapses was recently proposed [29]. Figure 8(a) shows how such DPI circuits can be integrated with our proposed spintronic synapses to generate exponential postsynaptic currents in response to input spikes. Assuming all transistors are in subthreshold saturation and using the translinear principle [28,29], it can be shown that the output current I_{syn} exhibits temporal dynamics of the form

$$\tau_{\mathrm{syn}} \frac{dI_{\mathrm{syn}}}{dt} + I_{\mathrm{syn}} = \frac{I_w I_{\mathrm{th}}}{I_t}, \tag{8}$$

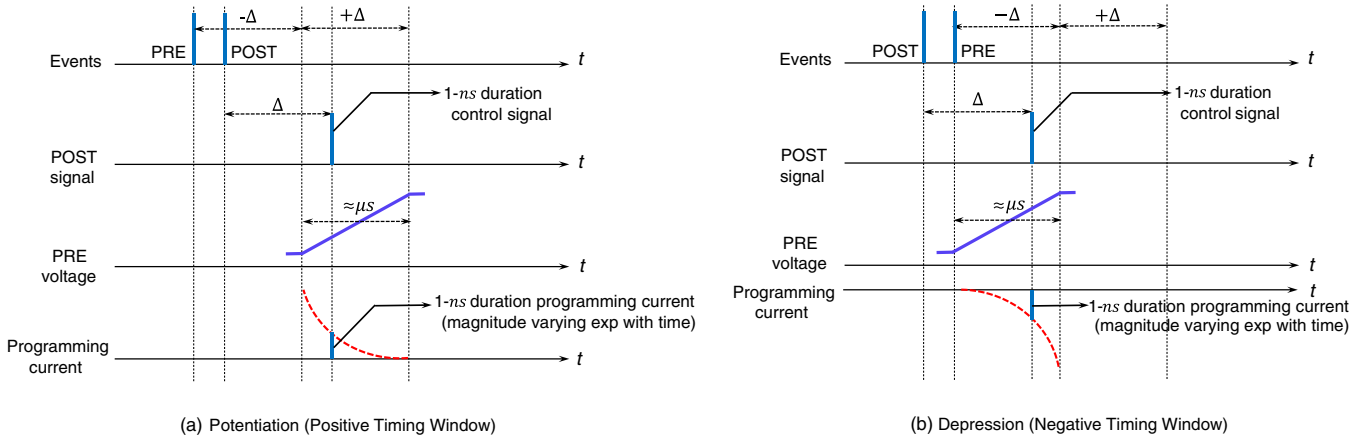

FIG. 7. Detailed timing diagrams demonstrating the implementation of (a) potentiation (positive timing window) and (b) depression (negative timing window) in the spintronic synapse. POST is the control signal that is activated during programing, while PRE is the gate voltage of the M_{STDP} transistor that implements synaptic plasticity. Duration of the programing current is determined by the duration of the POST signal, while the magnitude is determined by the value of the PRE signal when the POST signal is high.

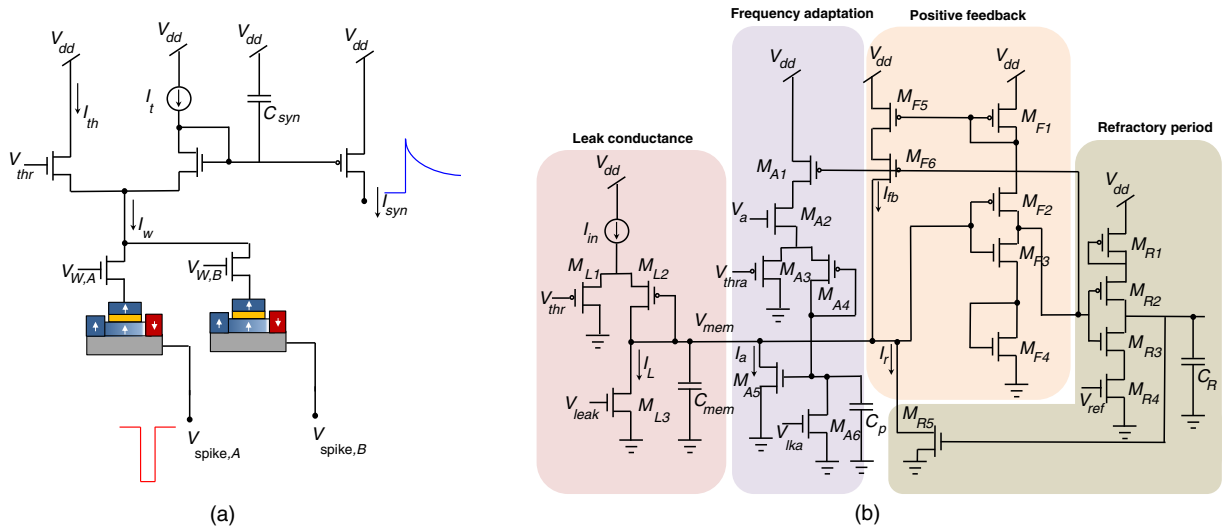

FIG. 8. (a) DPI circuit interfaced with spintronic synapses to emulate synaptic dynamics. (b) Subthreshold CMOS neuron with leak conductance, spike-frequency adaptation, positive feedback, and refractory-period implementation blocks [28].

where $\tau = (CU_T/kI_t)$. The above relationship is valid if the circuit is operated in the linear region ($I_t \ll I_w$). The bias voltage V_w acts as a scaling gain factor for the postsynaptic current. On the arrival of an input spike, the current I_w gets modulated by the MTJ conductance and thereby causes $I_{\rm syn}$ to increase by an amount governed by the synaptic weight. When there is no spike transmission, $I_{\rm syn}$ decreases exponentially, thereby emulating the synaptic dynamics discussed earlier. The access transistors driven by the $\overline{\rm POST}$ signal are not shown in Fig. 8 but are present in the design to ensure that the programing-current path is deactivated when the spike-transmission path is enabled.

C. Subthreshold CMOS neuron

CMOS circuits operating in subthreshold [Fig. 8(b)] have been shown to replicate a wide range of temporal dynamics observed in biological neurons like spike-frequency adaptation and refractory-period generation [28,30,31]. When operated in the subthreshold regime, the main mechanism of carrier transport in CMOS transistors is diffusion, thereby emulating the mechanism of ion flow in biological-neuron channels [28].

$I_{\rm in}$ represents the input current provided to the neuron. Using the translinear principle and assuming all transistors in subthreshold saturation, it can be shown that the temporal dynamics of $I_{\rm mem}$ is given by [28]

$$\tau_{\rm mem}\frac{dI_{\rm mem}}{dt} + I_{\rm mem}\left(1+\frac{I_a}{I_t}\right) = \frac{I_{\rm in}I_{\rm th}}{I_t}, \qquad (9)$$

where $\tau = (C_{\rm mem}U_T/kI_t)$. The above relation is again valid when the DPI circuit operates in the linear region (i.e., $I_t \ll I_{\rm in}$).

We would like to conclude this section by relating the computing models discussed in Sec. II to the circuit implementations discussed in Sec. IV. Postsynaptic and neuron dynamics [referred to in Eqs. (2) and (5)] can be directly mapped to the DPI circuit and the subthreshold CMOS neuron circuit [referred to in Eqs. (8) and (9)], respectively. Readers are referred to Ref. [28] for details on neuromorphic chips utilizing such analog CMOS neurons and interfacing such circuits with post-CMOS synaptic crossbar arrays. Our proposal in this work includes the implementation of plasticity mechanism [referred to in Eq. (4)] in the spintronic-device structure utilizing the device concepts (presented in Sec. III) and the learning-circuit primitives (presented in Sec. IV A).

V. SIMULATION RESULTS

A. Simulation framework

In order to simulate the SNN implementation based on the proposed spintronic synapse, a hierarchical simulation framework is utilized. Device-level simulations of the spin-orbit torque-induced domain-wall motion is performed in MUMAX [32], a graphics-processing-unit-accelerated micromagnetic simulation tool. A behavioral model of the device is developed for the subsequent simulation of such synapses interfaced with CMOS neurons and learning circuits. The circuit-level simulations are performed in HSPICE using a standard cell library in commercial 45-nm CMOS technology. The device and circuit simulations are utilized to generate models of the plastic synapses and spiking neurons to perform system-level simulations of a network of spiking neurons using the BRIAN simulator [33].

B. Device-level simulations

The magnetization dynamics of the ferromagnet can be described by solving the Landau-Lifshitz-Gilbert equation with an additional term to account for the spin-orbit torque

TABLE I. Device-simulation parameters.

Parameters	Value
Ferromagnet dimensions	$320 \times 20 \times 0.6$ nm^3
Grid size	$4 \times 1 \times 0.6$ nm^3
Heavy-metal thickness	3 nm
Domain-wall width	7.6 nm
Saturation magnetization, M_s	700 K A/m
Spin Hall angle, θ	0.07
Gilbert-damping factor, α	0.3
Exchange-correlation constant, A	1×10^{-11} J/m
Perpendicular magnetic anisotropy	4.8×10^5 J/m^3
Effective DMI constant, D	-1.2×10^{-3} J/m^2

generated by the spin Hall effect at the FM-HM interface [21,34],

$$\frac{d\hat{\mathbf{m}}}{dt} = -\gamma(\hat{\mathbf{m}} \times \mathbf{H}_{\text{eff}}) + \alpha\left(\hat{\mathbf{m}} \times \frac{d\hat{\mathbf{m}}}{dt}\right) + \beta(\hat{\mathbf{m}} \times \hat{\mathbf{m}}_P \times \hat{\mathbf{m}}), \tag{10}$$

where $\hat{\mathbf{m}}$ is the unit vector of the FM magnetization at each grid point, $\gamma = (2\mu_B\mu_0/\hbar)$ is the gyromagnetic ratio for electron, α is Gilbert's damping ratio, $\mathbf{H}_{\text{eff}}$ is the effective magnetic field, $\beta = (\hbar\theta J/2\mu_0 etM_s)$ [where $\hbar$ is Planck's constant, J is the input charge current density, θ is the spin Hall angle [21], μ_0 is the permeability of the vacuum, e is the electronic charge, t is the FM thickness, and M_s is the saturation magnetization], and $\hat{\mathbf{m}}_P$ is the direction of the input spin current. The effective field $\mathbf{H}_{\text{eff}}$ also includes the field resulting from the DMI and is given by

$$\mathbf{H}_{\text{DMI}} = -\frac{2D}{\mu_0 M_s}\left[\frac{\partial m_z}{\partial x}\hat{x} + \frac{\partial m_z}{\partial y}\hat{y} - \left(\frac{\partial m_x}{\partial x} + \frac{\partial m_y}{\partial y}\right)\hat{z}\right]. \tag{11}$$

Here, D represents the effective DMI constant and determines the strength of the DMI field in such multilayer structures. A positive sign of D implies right-handed chirality, and vice versa. In the presence of DMI, the boundary conditions at the edges of the sample are given by

$$\frac{\partial\hat{\mathbf{m}}}{\partial n} = \frac{D}{2A}\hat{\mathbf{m}} \times (\hat{\mathbf{n}} \times \hat{z}), \tag{12}$$

where A is the exchange correlation constant and $\hat{\mathbf{n}}$ represents the unit vector normal to the surface of the FM. The simulation parameters are given in Table I and are used for the rest of this work, unless otherwise stated. The parameters are obtained experimentally from magnetometric measurements of Ta(3 nm)/Pt(3 nm)/CoFe(0.6 nm)/MgO(1.8 nm)/Ta(2 nm) nanostrips [22]. Current density is estimated by assuming that the current flow is mainly through the FM-HM layers in the stack structure [22].

Figure 9(a) shows the domain-wall displacement in a CoFe sample with a cross section of 160×0.6 nm for a charge-current density of $J = 0.1 \times 10^{12}$ A/m^2. The grid size is taken to be $4 \times 4 \times 0.6$ nm^3. Figure 9(b) depicts the variation of the domain-wall velocity with the input charge-current density. The velocity increases linearly with the current density and ultimately reaches a saturation velocity. The graphs are in good agreement with the results illustrated in Ref. [21] for the same multilayer structure described in this section. Figure 9(c) illustrates the fact that the domain-wall displacement is directly proportional to the magnitude of the programing current (for domain-wall velocities below the saturation regime). For a duration of 1 ns, a maximum current of approximately 80 μA is required to displace the domain wall from one edge of the FM to the other.

A nonequilibrium-Green's-function- (NEGF-)based transport-simulation framework [35] is used to model the variation of the MTJ resistance with the oxide thickness [Fig. 10(a)] and the applied voltage [Fig. 10(b)], respectively. In order to determine the MTJ resistance for a FM with a domain wall separating two oppositely polarized magnetized domains, the NEGF-based simulator [35] is modified by considering the parallel connection of three MTJs. The magnetization directions of the FLs of the three MTJs are considered parallel, antiparallel, and perpendicular (domain wall) to the pinned-layer magnetization. The length of the first two MTJs is varied according to the position of the domain wall, while the width of the third MTJ is taken to be equal to the domain-wall width. Figure 11(a) depicts the

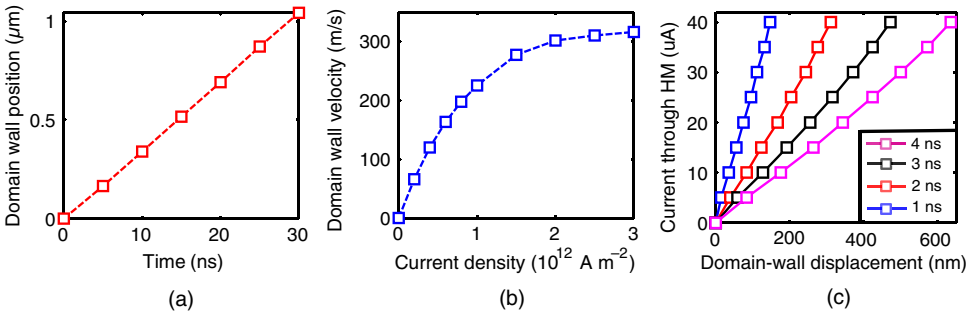

FIG. 9. (a) Domain-wall displacement as a function of time for a CoFe strip of cross section 160×0.6 nm due to the application of a charge-current density, $J = 0.1 \times 10^{12}$ A/m^{-2}. (b) Domain-wall velocity as a function of current density. The results are in good agreement with Ref. [21]. (c) Domain-wall displacement is directly proportional to the programing current for a fixed duration of the programing pulse.

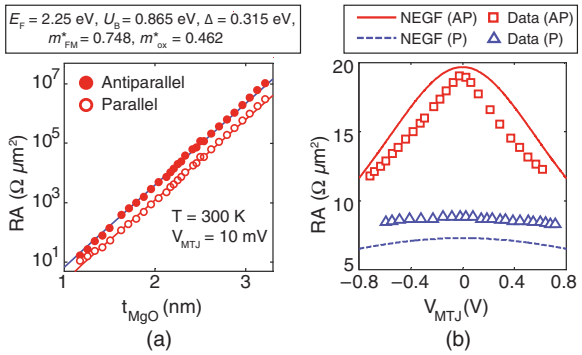

FIG. 10. The NEGF-based transport-simulation framework is calibrated to the experimental results illustrated in Refs. [36,37]. MTJ resistance varies with (a) oxide thickness and (b) applied voltage.

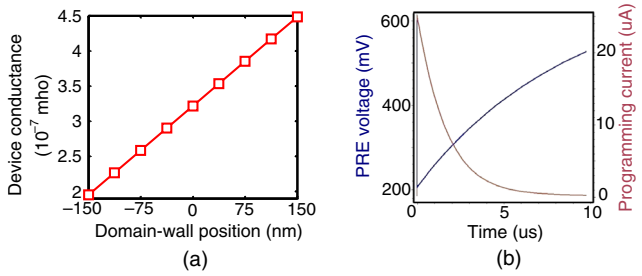

FIG. 11. (a) Linear variation of device conductance with domain-wall position. (b) Programing-circuit simulation to generate the STDP characteristics in the proposed spintronic synapse.

variation of the device conductance with the domain-wall position (with an origin at the middle of the FM). In order to ensure proper synaptic functionality, it is also essential that the device resistance (for a particular position of the domain wall) does not vary with the voltage drop across the device. This proper functionality is ensured by appropriately interfacing the device with the DPI circuit discussed earlier to generate the synaptic dynamics. The range of synapse resistances are in the MΩ range, while the current flowing through the MTJ is in the range of a few nanoamperes. Hence, the voltage drop across the MTJ should be approximately a few millivolts (< 100 mV). It is apparent from Fig. 10(b) that the operating range of V_{MTJ} is low enough to ensure a negligible variation of the device conductance against the device voltage drop for a particular domain-wall position. As explained previously, such a linear variation of the device conductance against the domain-wall position results in the programing current being directly proportional to the relative conductance (weight) change involved. Hence, the temporal profile of the necessary programing current also follows the STDP characteristics.

C. Circuit-level simulations

The programing and neuron circuits are simulated using a standard cell library in 45-nm commercial CMOS technology. Although biological time scales are in the range of milliseconds, it is not essential to limit the processing speed of the circuit to such slow time constants for implementing pattern-recognition systems [6]. The circuits are designed to operate at time constants in the range of microseconds.

Figure 11(b) shows the response of the programing circuit for the case where the programing-current path is active throughout the simulation time. The gate voltage of the transistor M_{STDP} increases linearly and is reset at each input pulse, leading to exponential subthreshold current dynamics. The average power consumption of the circuit is 0.46 μW for the entire positive time window. The duration of the time window can be varied by changing the capacitance value. Further, this programing circuit can be shared by synapses in a particular column. It is worth noting here that this power consumption does not include the power consumed in the M_{STDP} transistor, as current will flow through it only when the programing-current path is activated for 1 ns. The supply voltage for M_{STDP} transistor is maintained at 600 mV; hence, the maximum amount of energy consumption involved in synapse programing is approximately 48 fJ (600 mV × 80 μA × 1 ns) per synaptic event.

Figure 12 depicts the response of the CMOS neuron to a constant input current. As explained earlier, the spike-frequency adaptation scheme reduces the spike frequency to a steady-state value. For a membrane capacitance of 50 fF, the average power consumption of the circuit is approximately 5.7 pJ/spike.

D. System-level simulations

The device and circuit behavioral models are used to simulate a SNN for digit-recognition problems. The input images (28 × 28 pixels) used for training is taken from the MNIST data set [38]. The images are rate encoded, and an

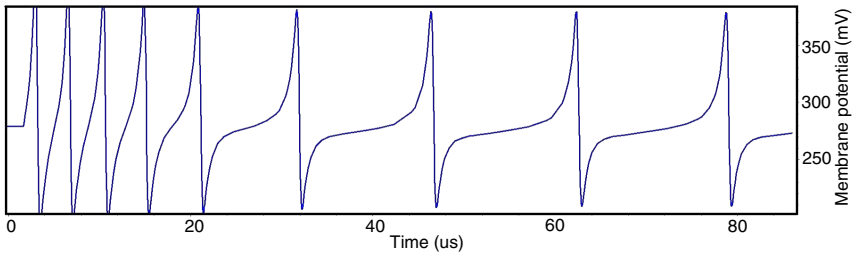

FIG. 12. CMOS neuron response to a constant input current with positive feedback, spike-frequency adaptation, and refractory-period implementations.

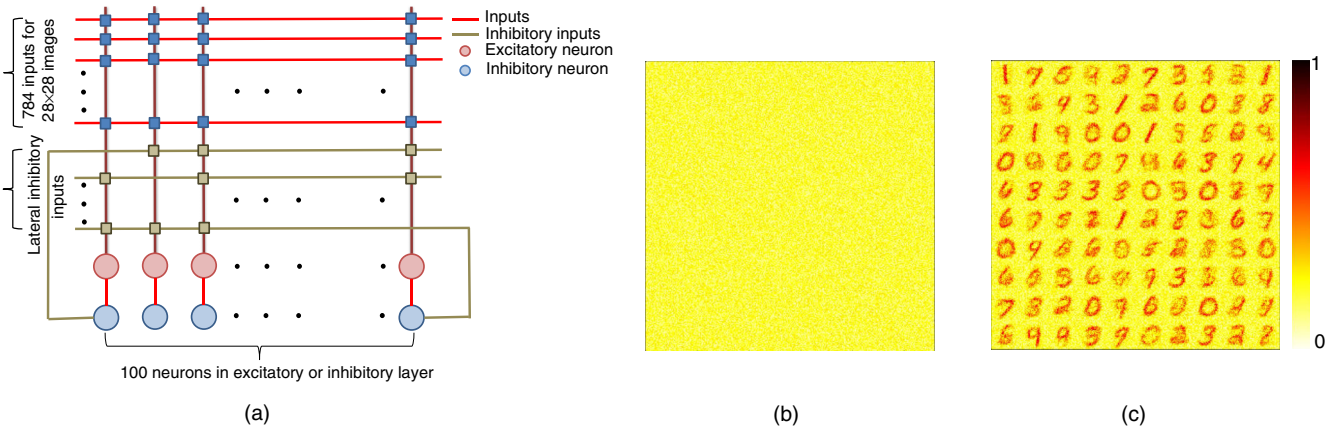

FIG. 13. (a) SNN topology used for digit recognition arranged in a crossbar array. (b) Initial random synapse weights plotted in a 28×28 array for 100 neurons in the excitatory layer. (c) Representative digit patterns start getting stored in the synapse weights for each neuron after 1000 learning epochs.

array of 100 excitatory neurons is used to simulate the self-learning functionality of synapses in SNNs. Figure 13(a) demonstrates the SNN topology used for the recognition problem arranged in a crossbar array. Synapses present at the cross points joining the inputs to the excitatory neurons can be programed depending on the temporal spiking patterns of the pre- and postneurons. Note that a synapse is absent at the cross point joining the excitatory to the inhibitory neuron. Inhibitory neurons are exactly similar to the excitatory neurons except that the output voltage spikes are negative.

Figures 13(b) and 13(c) depict synapse weights plotted in a 28×28 array (same as for the input images) for each of the 100 neurons used for recognition purposes. Initially, all of the weights are random. However, as learning progresses, the synapses of each neuron start learning generic representations of the various digits. Thus, a particular neuron becomes more sensitive to the digit whose generic representation is being stored in its synapse weights since it will fire more if input spike trains are received at the pixel locations corresponding to high synaptic weights. The various system-level simulation parameters are outlined in Table II. The parameters are tuned to achieve learning ability in the synapses. The units of the time constants are given with respect to the duration of each time step in the simulation. For this work, the circuits are designed to operate in a microsecond time scale, as mentioned before. It is worth noting here that the manner in which the time constants and the other parameters can be tuned in the circuit-level simulations were discussed in Sec. IV. Each number in parentheses represents the value corresponding to the inhibitory neuron.

TABLE II. System simulation parameters.

Parameters	Value
No. of excitatory or inhibitory neurons	100
Probability of input spike per time step	0–0.06375
Number of time steps per image	350
STDP time constants	100 (1)
Neuron time constants	10 (10)
Postsynaptic-current time constants	1 (2)

Additionally, we would like to mention here that such neuromorphic systems are significantly robust to imprecision due to device mismatch, variability, and noise effects owing to the adaptive nature of such computations involving plasticity, homeostasis, and feedback mechanisms [28]. Furthermore, Querlioz *et al.* [39] demonstrated the immunity of such single-layer SNNs based on crossbar arrays of resistive synapses with lateral-inhibition and homeostasis effects to variations and nonidealities in typical resistive synaptic devices and CMOS neuron circuits. Specifically, we perform an analysis of the impact of variations in the oxide thickness or equivalently the MTJ synaptic conductances on the classification accuracy of the system. Almost no degradation in classification accuracy is observed for the 100-neuron network, even with a 25% variation in the resistances of the spintronic synapses.

VI. CONCLUSIONS

Prior proposals have investigated monodomain spintronic devices for implementing spiking neurons [40] and short-term plasticity effects [41]. Here we propose a hybrid spintronic-CMOS SNN design with self-learning (from the device to the system level) based on a three-terminal multidomain spintronic-synapse-device structure consisting of decoupled spike-transmission and programming-current paths. This design is advantageous for the implementation of neuromorphic systems capable of on-chip learning since the programing-current path is independent of the read-current path. Interface CMOS circuit design for self-learning is highly simplified since the

TABLE III. Comparison with other proposed synapses.

Device	Dimensions	Programing energy/ operating voltage	Programing time	Terminals	Programing mechanism
GeSbTe memristor [7]	40-nm mushroom and 10-nm pore	Average 2.74 pJ/event	60 ns	2	Programed by Joule heating (phase change)
GeSbTe memristor [11]	75-nm electrode diameter	50 pJ (reset) and 0.675 pJ (set)	10 ns	2	Programed by Joule heating (phase change)
Ag/AgInSbTe/Ag chalcogenide memristor [42]	100 μm × 100 μm	Threshold voltage: 0.3 V	5 μs	2	Programed by Joule heating (phase change)
Ag-Si memristor [8]	100 nm × 100 nm	Threshold voltage: 2.2 V	300 μs	2	Movement of Ag ions
FeFET [10]	Channel length: 3 μm	Maximum gate voltage: 4 V	10 μs	3	Gate-voltage modulation of ferroelectric polarization
Floating-gate transistor [9]	1.8 μm/0.6 μm (0.35 μm; CMOS technology)	$V_{dd} - 4.2$ V and tunneling voltage: 15 V	100 μs (injection) and 2 ms (tunneling)	3	Injection and tunneling currents
SRAM synapse [6]	0.3 μm^2 (10 nm; CMOS technology)	Average 328 fJ (4-bit synapse)	···	···	Digital-counter-based circuits
Spintronic synapse	Ferromagnet dimensions: 320 nm × 20 nm	Maximum 48 fJ/event	1 ns	3	Spin-orbit torque

resistance in the programing-current path is constant and determined mainly by the HM resistance and is independent of the synapse conductance.

Table III provides a comparative analysis of our spintronic synapse (calibrated to experiments performed in Ref. [22]) with other proposed synaptic devices. Synaptic device structures based on emerging post-CMOS technologies [7,8,11] are usually two-terminal devices and do not offer decoupled programing and read-current paths. Additionally, they are usually characterized by relatively high programing energies. In contrast, our proposed synapse offers low programing energy and requires very little programing time. A maximum programing energy of approximately 48 fJ is consumed per synaptic event due to the highly-energy-efficient spin-orbit torque-induced synaptic plasticity. Three terminal synaptic devices based on ferroelectric field-effect transistor (FeFET) [10] and floating-gate transistors [9] have also been proposed. However, the programing in such devices is usually accomplished through the gate terminal, and a high gate voltage is usually applied across a very thin oxide [9,10], leading to reliability issues, in addition to associated high power consumption. Programing is also relatively slow in such three-terminal synaptic devices [9,10]. It is worth noting here that the current flowing through the oxide in the MTJ structure for our proposed synapse is the read current, which is in the range of nanoamperes and drives subthreshold CMOS circuits. Static-random-access-memory-(SRAM-)based synapses have been also proposed for digital CMOS-based SNN designs [6]. However, for implementing 1 bit of the synapse, an 8-T SRAM cell has to be used, thereby leading to significant area overhead for implementation of a single synapse [6]. In addition, learning circuits involve multiple digital counters and are more area and power consuming than our proposed design.

Please see Ref. [43] for a discussion on the practical implementation of arrays of such spintronic devices interfaced with CMOS transistors. The size limitation of crossbar arrays of such spintronic devices is determined by the driving capabilities of rows of the array by input voltages in the presence of parasitics. In addition, sneak paths also become a potential issue for large crossbar arrays in order to implement on-chip learning. These concerns are equally valid, in general, for spin devices and other memristive technologies. However, it is worth noting here that computation occurring in a large crossbar can be distributed easily among smaller crossbar arrays by simply replacing the large unit by an equivalent number of smaller crossbar units using peripheral-control circuitry.

In conclusion, in this work, we formulate a device, circuit, and algorithm cosimulation framework calibrated to experimental results to validate the functionalities and the performance of the proposed hybrid spintronic-CMOS–based SNN design with on-chip learning. We propose circuit primitives for generating STDP in the proposed synapse and demonstrate how such synaptic devices could be arranged in a crossbar fashion leading to an area- and power-efficient SNN implementation that is capable of recognizing patterns in input data. Simulation studies indicate the efficiency of the proposed hybrid spintronic-CMOS–based SNN design as an ultralow-power neuromorphic-computing platform capable of online learning.

ACKNOWLEDGMENTS

This work was supported, in part, by the Center for Spintronic Materials, Interfaces, and Novel Architectures (C-SPIN), a MARCO- and DARPA-sponsored StarNet center, the Semiconductor Research Corporation, the National Science Foundation, the Intel Corporation, and the National Security Science and Engineering Faculty Fellowship.

[1] S. Ghosh-Dastidar and H. Adeli, Spiking neural networks, International Journal of Neural Systems **19**, 295 (2009).
[2] H. Markram, The blue brain project, Nat. Rev. Neurosci. **7**, 153 (2006).
[3] J. Schemmel, J. Fieres, and K. Meier, in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), Shatin, Hong Kong, 2008* (IEEE, New York, 2008), p. 431.
[4] X. Jin, M. Lujan, L. A. Plana, S. Davies, S. Temple, and S. Furber, Modeling spiking neural networks on SpiNNaker, Comput. Sci. Eng. **12**, 91 (2010).
[5] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, A million spiking-neuron integrated circuit with a scalable communication network and interface, Science **345**, 668 (2014).
[6] B. Rajendran, Y. Liu, J.-s. Seo, K. Gopalakrishnan, L. Chang, D. J. Friedman, and M. B. Ritter, Specifications of nanoscale devices and circuits for neuromorphic computational systems, IEEE Trans. Electron Devices **60**, 246 (2013).
[7] B. L. Jackson, B. Rajendran, G. S. Corrado, M. Breitwisch, G. W. Burr, R. Cheek, K. Gopalakrishnan, S. Raoux, C. T. Rettner, A. Padilla *et al.*, Nanoscale electronic synapses using phase change devices, ACM J. Emerging Technol. Comput. Syst. **9**, 12 (2013).
[8] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, Nanoscale memristor device as synapse in neuromorphic systems, Nano Lett. **10**, 1297 (2010).
[9] S. Ramakrishnan, P. E. Hasler, and C. Gordon, Floating gate synapses with spike-time-dependent plasticity, IEEE Trans. Biomed. Circuits Syst. **5**, 244 (2011).
[10] Y. Nishitani, Y. Kaneko, M. Ueda, E. Fujii, and A. Tsujimura, Dynamic observation of brain-like learning in a ferroelectric synapse device, Jpn. J. Appl. Phys. **52**, 04CE06 (2013).
[11] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing, Nano Lett. **12**, 2179 (2012).
[12] M. Sharad, C. Augustine, G. Panagopoulos, and K. Roy, Spin-based neuron model with domain-wall magnets as synapse, IEEE Trans. Nanotechnol. **11**, 843 (2012).
[13] S. G. Ramasubramanian, R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, in *Proceedings of the International Symposium on Low Power Electronics and Design, La Jolla, 2014* (ACM, New York, 2014), p. 15.
[14] A. Sengupta, S. H. Choday, Y. Kim, and K. Roy, Spin orbit torque based electronic neuron, Appl. Phys. Lett. **106**, 143701 (2015).
[15] R. G. Morris, D. O. Hebb: The organization of behavior, Brain Research Bulletin **50**, 437 (1999).
[16] G.-q. Bi and M.-m. Poo, Synaptic modification by correlated activity: Hebb's postulate revisited, Annu. Rev. Neurosci. **24**, 139 (2001).
[17] P. U. Diehl and M. Cook, Unsupervised learning of digit recognition using spike-timing-dependent plasticity, Front. Comput. Neurosci. **9**, 99 (2015).
[18] P. Knag, J. K. Kim, T. Chen, and Z. Zhang, A sparse coding neural network ASIC with on-chip learning for feature extraction and encoding, IEEE J. Solid-State Circuits **50**, 1070 (2015).
[19] S. Emori, U. Bauer, S.-M. Ahn, E. Martinez, and G. S. Beach, Current-driven dynamics of chiral ferromagnetic domain walls, Nat. Mater. **12**, 611 (2013).
[20] G. Chen, J. Zhu, A. Quesada, J. Li, A. N'Diaye, Y. Huo, T. Ma, Y. Chen, H. Kwon, C. Won *et al.*, Novel Chiral Magnetic Domain Wall Structure in Fe/Ni/Cu(001) Films, Phys. Rev. Lett. **110**, 177204 (2013).
[21] E. Martinez, S. Emori, N. Perez, L. Torres, and G. S. Beach, Current-driven dynamics of Dzyaloshinskii domain walls in the presence of in-plane fields: Full micromagnetic and one-dimensional analysis, J. Appl. Phys. **115**, 213909 (2014).
[22] S. Emori, E. Martinez, K.-J. Lee, H.-W. Lee, U. Bauer, S.-M. Ahn, P. Agrawal, D. C. Bono, and G. S. Beach, Spin Hall torque magnetometry of Dzyaloshinskii domain walls, Phys. Rev. B **90**, 184427 (2014).
[23] N. Perez, L. Torres, and E. Martinez-Vecino, Micromagnetic modeling of Dzyaloshinskii-Moriya interaction in spin Hall effect switching, IEEE Trans. Magn. **50**, 1 (2014).
[24] J. E. Hirsch, Spin Hall Effect, Phys. Rev. Lett. **83**, 1834 (1999).
[25] A. Sengupta, Z. Al Azim, X. Fong, and K. Roy, Spin-orbit torque induced spike-timing dependent plasticity, Appl. Phys. Lett. **106**, 093704 (2015).
[26] J. Lazzaro and J. Wawrzynek, *Low-Power Silicon Neurons, Axons and Synapses* (Springer, New York, 1994).
[27] C. Bartolozzi and G. Indiveri, Synaptic dynamics in analog VLSI, Neural Comput. **19**, 2581 (2007).
[28] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, Neuromorphic electronic circuits for building autonomous cognitive systems, Proc. IEEE **102**, 1367 (2014).
[29] G. Indiveri, B. Linares-Barranco, R. Legenstein, G. Deligeorgis, and T. Prodromakis, Integration of nanoscale memristor synapses in neuromorphic computing architectures, Nanotechnology **24**, 384010 (2013).
[30] G. Indiveri, in *IEEE International Symposium on Circuits and Systems (ISCAS), Bangkok, 2003* (IEEE, New York, 2003), p. 820.
[31] P. Livi and G. Indiveri, in *IEEE International Symposium on Circuits and Systems (ISCAS), Taipei, 2009* (IEEE, New York, 2009), p. 2898.
[32] A. Vansteenkiste, J. Leliaert, M. Dvornik, M. Helsen, F. Garcia-Sanchez, and B. Van Waeyenberge, The design and verification of MUMAX3, AIP Adv. **4**, 107133 (2014).
[33] D. F. Goodman and R. Brette, The BRIAN simulator, Front. Neurosci. **3**, 192 (2009).

[34] J. C. Slonczewski, Conductance and exchange coupling of two ferromagnets separated by a tunneling barrier, Phys. Rev. B **39**, 6995 (1989).

[35] X. Fong, S. K. Gupta, N. N. Mojumder, S. H. Choday, C. Augustine, and K. Roy, in *Proceedings of the International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), Osaka, Japan, 2011* (IEEE, New York, 2011), p. 51.

[36] S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando, Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions, Nat. Mater. **3**, 868 (2004).

[37] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu *et al.*, in *Proceedings of the IEEE International Electron Devices Meeting (IEDM), Baltimore, 2009* (IEEE, New York, 2009), p. 1.

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[39] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, Immunity to device variations in a spiking neural network with memristive nanodevices, IEEE Trans. Nanotechnol. **12**, 288 (2013).

[40] A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, and K. Roy, Magnetic tunnel junction mimics stochastic cortical spiking neurons, Sci. Rep. **6**, 30039 (2016).

[41] A. Sengupta and K. Roy, Short-Term Plasticity and Long-Term Potentiation in Magnetic Tunnel Junctions: Towards Volatile Synapses, Phys. Rev. Applied **5**, 024012 (2016).

[42] Y. Li, Y. Zhong, J. Zhang, L. Xu, Q. Wang, H. Sun, H. Tong, X. Cheng, and X. Miao, Activity-dependent synaptic plasticity of a chalcogenide electronic synapse for neuromorphic systems, Sci. Rep. **4**, 4906 (2014).

[43] H. Noguchi, K. Ikegami, K. Kushida, K. Abe, S. Itai, S. Takaya, N. Shimomura, J. Ito, A. Kawasumi, H. Hara *et al.*, in *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, 2015* (IEEE, New York, 2015), p. 1.